Course Title: **Artificial Intelligence and Expert Systems** Course Code: CCE3219 3rd year
Date: 18.4.2015 (Second term) Allowed time: 1 hrs

Answer the following questions:

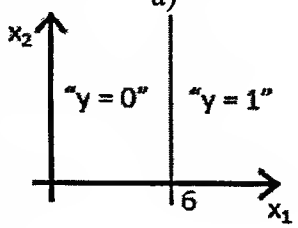
Question No. 1 **(8 marks)**

For each of the following, please circle the letter introducing the best answer. (Check all that apply.) Explain your answer.

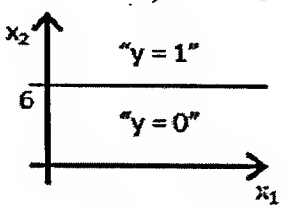
1. Suppose you are working on stock market prediction, and you would like to predict whether or not a particular stock's price will be higher tomorrow than it is today. You want to use a learning algorithm. Which one of the following algorithms is appropriate?
   a) Regression
   b) Classification
   c) Clustering
   d) Reinforcement learning
2. A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, what is T?
   a) The process of the algorithm examining a large amount of historical weather data.
   b) The weather prediction task.
   c) The probability of it correctly predicting a future date's weather.
   d) None of these.
3. Let f be some function so that $f(\theta_0, \theta_1)$ outputs a number. For this problem, f is some arbitrary/unknown smooth function (not necessarily the cost function of linear regression, so may have local optima). Suppose we use gradient descent to try to minimize $f(\theta_0, \theta_1)$ as a function of θ_0 and θ_1. Which of the following statements are true? (Check all that apply.)
   a) If the first few iterations of gradient descent cause $f(\theta_0, \theta_1)$ to increase rather than decrease, then the most likely cause is that we have set the learning rate α to too large a value.
   b) If θ_0 and θ_1 are initialized at a local minimum, the one iteration will not change their values.
   c) No matter how θ_0 and θ_1 are initialized, so long as α is sufficiently small, we can safely expect gradient descent to converge to the same solution.
   d) Setting the learning rate α to be very small is not harmful, and can only speed up the convergence of gradient descent.

4. Suppose you have a dataset with m=100000 examples and n=15 features for each example. You want to use multivariate linear regression to fit the parameters to our data. Should you prefer gradient descent or the normal equation?
   a) The normal equation, since gradient descent might be unable to find the optimal θ.
   b) The normal equation, since it provides an efficient way to directly find the solution.
   c) Gradient descent, since it will always converge to the optimal θ.
   d) Gradient descent, since $(X^TX)^{-1}$ will be very slow to compute in the normal equation.
5. Suppose you train a logistic classifier $h_\theta(x)=g(\theta_0+\theta_1x_1+\theta_2x_2)$. Suppose $\theta_0=-6$, $\theta_1=0$, *and* $\theta_2=1$. Which of the following figures represents the decision boundary found by your classifier?

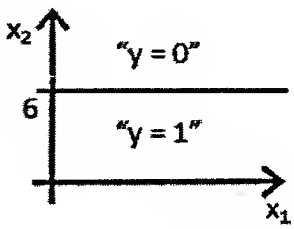
a)



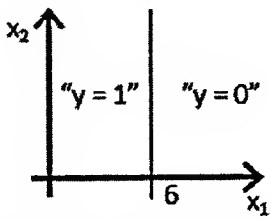
b)



c)



d)

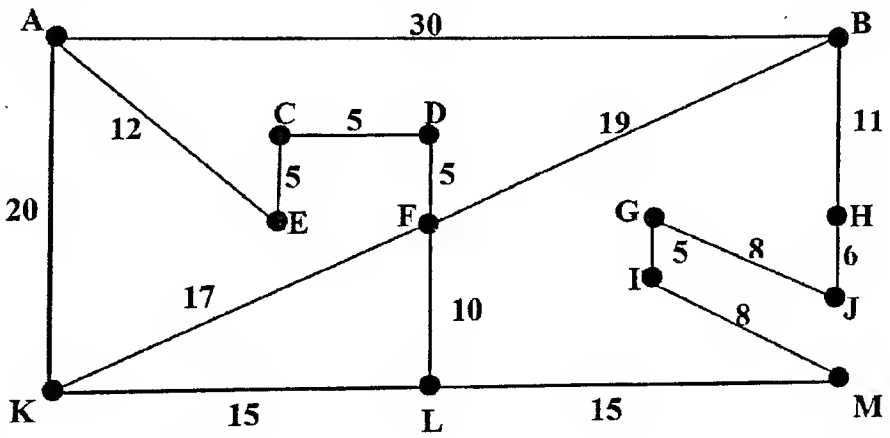


6. You are training a classification model with logistic regression. Which of the following statements are true? Check all that apply.
   a) Introducing regularization to the model always results in equal or better performance on the training set.
   b) Adding a new feature to the model always results in equal or better performance on the training set.
   c) Adding a new feature to the model always results in equal or better performance on examples not in the training set.
   d) Introducing regularization to the model always results in equal or better performance on examples not in the training set.
7. Consider an A* search algorithm for which h(n) = 0. To which of the following search algorithms is this A* equivalent?
   a) Greedy best-first search
   b) Depth-First Search
   c) Uniform Cost Search
   d) None of the above.

8. K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?
   a) The cluster assignment step, where the parameters $C^{(i)}$ are updated.
   b) Move the cluster centroids, where the centroids μ_k are updated.
   c) Feature scaling, to ensure each feature is on a comparable scale to the others.
   d) Using the elbow method to choose K.

Question No. 2 **(12 marks)**

1. Consider the following map (not drawn to scale).



Use the A* algorithm to work out a route from town A to town M. Use the following cost functions.

- G(n) = The cost of each move as the distance between each town (shown on map).
- H(n) = The Straight Line Distance between any town and town M. These distances are given in the table below.

Straight Line Distance to M

A	B	C	D	E	F	G	H	I	J	K	L	M
56	22	30	29	29	30	14	10	8	5	30	15	0

   a) Provide the search tree for your solution, showing the order in which the nodes were expanded and the cost at each node. You should not re-visit a town that you have just come from. State the route you would take and the cost of that route.
   b) Assume the estimated costs by the heuristic were replaced and shown in the following table

Straight Line Distance to M

A	B	C	D	E	F	G	H	I	J	K	L	M
80	10	50	20	10	30	60	30	20	50	60	20	0

What route would now be returned by the A* algorithm and what would the cost of that route be?

   c) Comment on the optimality of the two A* algorithms. How do you account for the different routes returned?

2. Prove that A* tree search with admissible heuristic is optimal.

3. Consider the problem of predicting how well a student does in his second year of college/university, given how well they did in their first year. Specifically, let x be equal to the number of "A" grades (including A-. A and A+ grades) that a student receives in their first year of college (freshmen year). We would like to predict the value of y, which we define as the number of "A" grades they get in their second year.
Use the following training set of a small sample of different students' performances. Here each row is one training example.

x	y
3	2
1	2
0	1
4	3
5	4
3	4

Recall that in linear regression, the hypothesis is

$$h_\theta(x) = \theta_0 + \theta_1 x$$

, and we use m to denote the number of training examples.

a) For the training set given above, what is the value of m?
b) Recall the definition of the cost function What is $J(0, 1)$?

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

c) Suppose we set $\theta_0=-1$ and $\theta_1=2$. What is $h_\theta(6)$?

Best wishes

Dr. Sherin El Gokhy